Kaggle
鐵達尼號生存預測

10636006 錢佩青
10636008 游期順
10636026 陳亞蔚

目錄

一、 摘要

本組參加 Kaggle 中鐵達尼號生存預測競賽，目標透過乘客的資訊去預估這個乘客是否會在鐵達尼號沈船的意外中生存下來。

二、 介紹(研究背景及研究目的)

(1) 研究背景

鐵達尼號沉沒事故是 1912 年 4 月 14 日深夜至 15 日凌晨在北大西洋發生的著名船難，該船當時是世界最大的郵輪。當瞭望員看到冰山時，該船的行駛速度正接近最高速。由於無法快速轉向，該船右舷側面遭受了一次撞擊，部分船體出現縫隙，使 16 個水密隔艙中的 5 個進水。鐵達尼號的設計僅能夠承受 4 個水密隔艙進水，因此沉沒。

(2) 研究目的

目標乘客的資訊像是乘客的性別、姓名、出發港口、住的艙等、房間號碼、年齡、兄弟姊妹、老婆丈夫數量、父母小孩的數量、票的費用、票的號碼這些資訊去預估資料中這些乘客是否能夠在鐵達尼號沈船的意外中生存下來。

三、 資料及介紹(含資料特徵)及資料集來源

資料來源

Kaggle 網站: https://www.kaggle.com/c/titanic/data

資料介紹

- 訓練集資料筆數:891 筆
- 資料包含

  編號、是否存活、住的艙等、乘客的姓名、性別、年齡、兄弟姊妹＋老婆丈夫數量、父母小孩的數量、票的號碼、票的費用、房間號碼

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

四、 資料預處理

本組先透過資料分析後再決定要預處理的欄位，下面將先說明分析的項目後再說明將資料在放進模型前做的處理，包含欄位拆解、補空值..等。

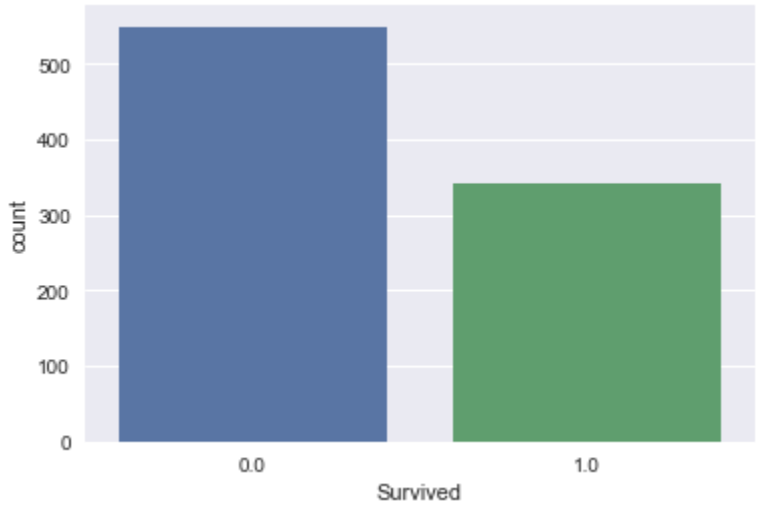
1. 資料分析

(1) 整體

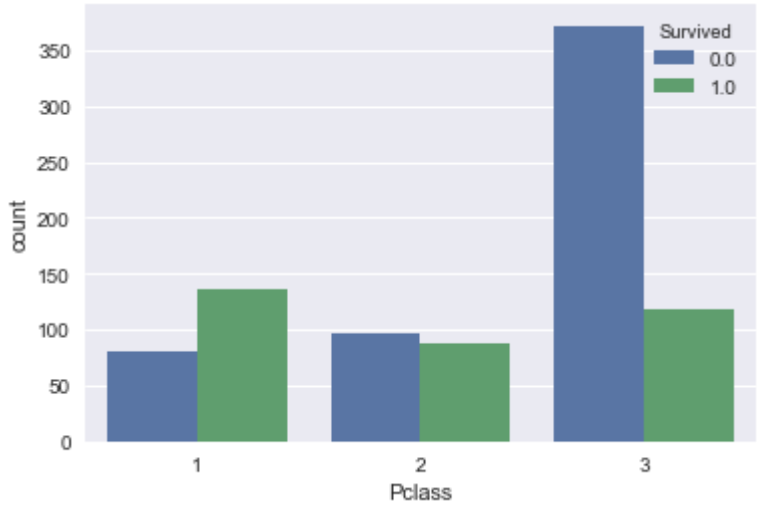
Train 資料的 Age,Cabin,Embark 欄位有空值以及 Test 資料的 Age, Fare, cabin 有空值

(2) 生存率

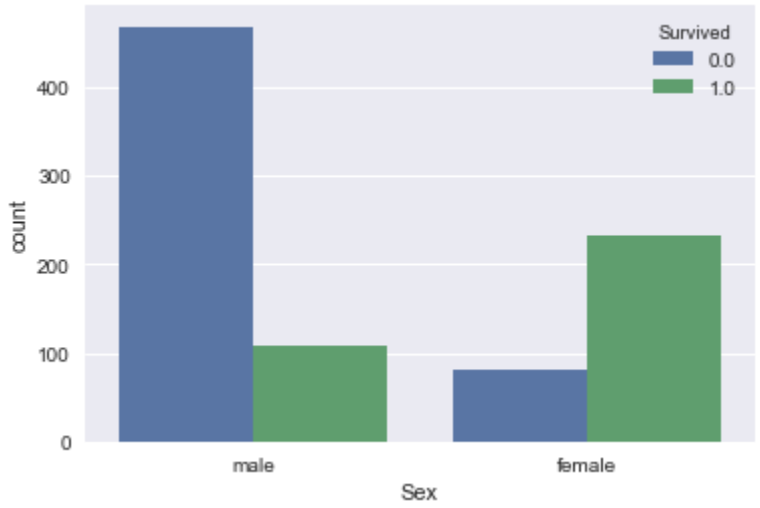
死亡的比例是 6 成、生存的比例大概是 4 成。



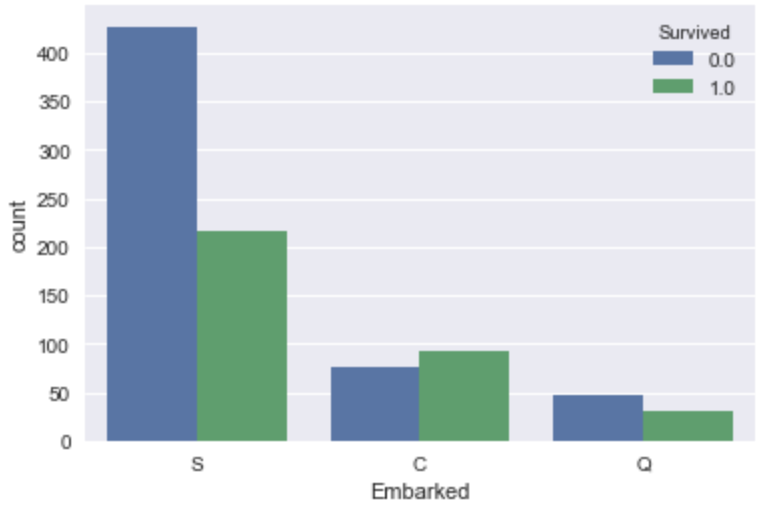
艙等跟生存率，1 艙等的生存率最高、再來是 2 艙等、最後是 3 艙等的。



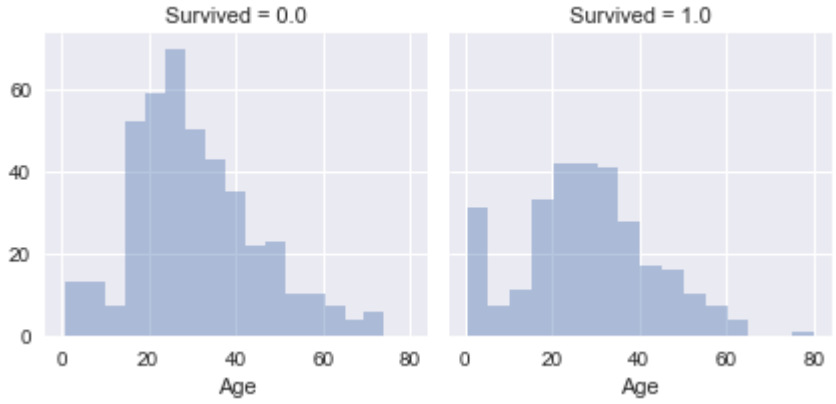
性別與生存率，女性生存率較男性高



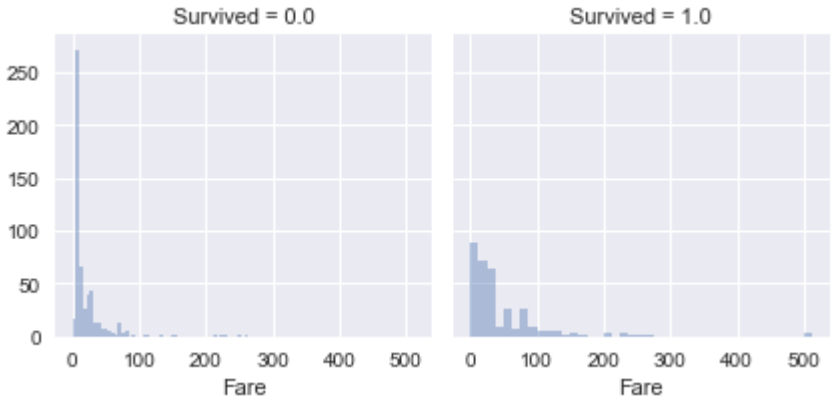
出發港口與生存率，S 港出發的生存率最低(原因也可能是 S 港的人買的票價都比較便宜)



年齡與生存率，年齡小的存活率較高。



票價與死亡率，票價低的乘客死亡率較高。



家庭大小與生存率，因資料中將父母＋小孩及兄弟姊妹＋丈夫妻子分成 2 欄位，因此將此兩欄位合併成家庭大小。

2. 資料預處理

(1) 姓名欄位

資料有稱謂的資訊(Mr., Miss.) ，稱謂總共有'Mr', 'Mrs','Miss', 'Master', 'Don', 'Rev', 'Dr', 'Mme', 'Ms', 'Major', 'Lady', 'Sir', 'Mlle', 'Col', 'Capt', 'the Countess', 'Jonkheer', 'Dona'

能計算出此些稱謂的平均年齡:

Capt	70.000000
Col	54.000000
Don	40.000000
Dona	39.000000
Dr	43.571429
Jonkheer	38.000000
Lady	48.000000
Major	48.500000
Master	5.482642
Miss	21.774238
Mlle	24.000000
Mme	24.000000
Mr	32.252151
Mrs	36.994118
Ms	28.000000
Rev	41.250000
Sir	49.000000
the Countess	33.000000

有部分稱謂的乘客是很少數的，因此將稱謂獨立出來以外也將部分較少的稱謂併入其他稱謂中，僅留 Master、Miss、Mr、Mrs。

(2) 票號欄位

票號的資訊取出前面英文的部分，相同的英文代碼可能代表的是房間的位置，後面的號碼不明白意義因此去除。

(3) 登船港口欄位

登船港口只遺漏少部分，因此補上最多數的 S 港。

(4) 費用欄位

補上平均值。

(5) 將類別資料轉為整數

五、 機器學習及深度學習方法

(1) 使用隨機森林來推測年齡

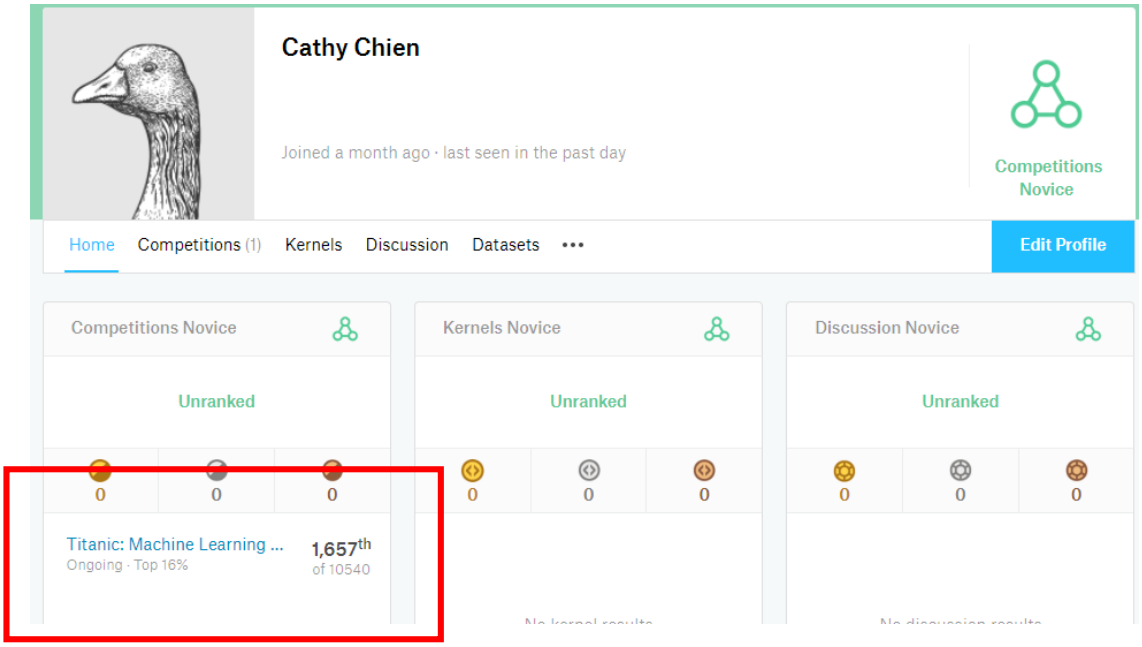
(2) 載入隨機森林演算法(Random Forest)來預測存活率

六、 研究結果及討論(含模型評估與改善)

透過研究發現，生存率與年齡、性別、票價、艙等…等等皆有相關，其中也能發現，女性及年紀小的乘客生存率皆較高。鐵達尼號電影中曾演出在搭乘救生艇時男性皆會讓女性、小孩先乘坐，依數據來看也與電影呈現吻合，可推測當時的確是此種狀況。

在本次中，僅用了隨機森林來推測年齡及存活率，若想再提升在 Kaggle 上的成績日後考慮可再多用幾種演算法像是投票法、裝袋法..等等。

Kaggle 排名如下圖



七、 結論

透過本次 Kaggle 競賽，讓我們能實際實作 Python 程式，對於上課所學習到的演算法也能再更加熟練地運用。在本次的鐵達尼號預測中也能夠透過分析了解鐵達尼號當時遇難時的逃難狀況，也在查閱參考資料中能夠更深入學習到深度學習及許多演算法的細節。希望在期中就能有這樣的機會能夠練習，並與班上其他同學討論分享，相信能學習到更多其他人的經驗。

附錄、參考文獻

(1) 鐵達尼號
https://zh.wikipedia.org/wiki/%E6%B3%B0%E5%9D%A6%E5%B0%BC%E5%85%8B%E5%8F%B7

(2) **Exploring Survival on the Titanic**
https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic

(3) 一個實例告訴你：**Kaggle** 數據競賽都有哪些套路
https://hk.saowen.com/a/68fa93708a2582abbf8851b6c407e689bae36ce2a20c3a1f64339f0f8a7ec009

(4) **Kaggle** 神器 **xgboost**
https://blog.csdn.net/aliceyangxi1987/article/details/72969146

(5) **[機器學習專案] Kaggle** 競賽**-**鐵達尼號生存預測
https://medium.com/@yulongtsai/https-medium-com-yulongtsai-titanic-top3-8e64741cc11f

(6) **[**資料分析**&**機器學習**]**第 **4.1** 講**:Kaggle** 競賽**-**鐵達尼號生存預測
https://medium.com/@yehjames/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E7%AC%AC4-1%E8%AC%9B-kaggle%E7%AB%B6%E8%B3%BD-%E9%90%B5%E9%81%94%E5%B0%BC%E8%99%9F%E7%94%9F%E5%AD%98%E9%A0%90%E6%B8%AC-%E5%89%8D16-%E6%8E%92%E5%90%8D-a8842fea7077